Theoretical Population Biology 107 (2016) 31–38

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Relationship inference from the genetic data on parents or offspring: A comparative study

Steven Gazal [a,b,c], Emmanuelle Génin [d,e,f], Anne-Louise Leutenegger [g,h,*]

[a] Inserm, UMR 1137, IAME, Paris, France
[b] Université Paris Diderot, Sorbonne Paris Cité, UMR 1137, Paris, France
[c] Plateforme de Génétique constitutionnelle-Nord (PfGC-Nord), Paris, France
[d] Inserm, UMR 1078, Brest, France
[e] Université Bretagne Occidentale, Brest, France
[f] Centre Hospitalier Régional Universitaire, Brest, France
[g] Inserm, U946, Genetic Variation and Human Diseases Lab, Paris, France
[h] Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, UMR 946, Paris, France

ARTICLE INFO

Article history:
Received 3 March 2015
Available online 30 September 2015

Keywords:
Relationship inference
Mating habit
Identity-by-descent
Kinship coefficient
Inbreeding coefficient
Genome sharing

ABSTRACT

Relationship inference in a population is of interest for many areas of research from anthropology to genetics. It is possible to directly infer the relationship between the two individuals in a couple from their genetic data or to indirectly infer it from the genetic data of one of their offspring. For this reason, one can wonder if it is more advantageous to sample couples or single individuals to study relationships of couples in a population. Indeed, sampling two individuals is more informative than sampling one as we are looking at four haplotypes instead of two, but it also doubles the cost of the study and is a more complex sampling scheme.

To answer this question, we performed simulations of 1000 trios from 10 different relationships using real human haplotypes to have realistic genome-wide genetic data. Then, we compared the genome sharing coefficients and the relationship inference obtained from either a pair of individuals or one of their offspring using both single-point and multi-point approaches.

We observed that for relationships closer than 1st cousin, pairs of individuals were more informative than one of their offspring for relationship inference, and kinship coefficients obtained from single-point methods gave more accurate or equivalent genome sharing estimations. For more remote relationships, offspring were more informative for relationship inference, and inbreeding coefficients obtained from multi-point methods gave more accurate genome sharing estimations.

In conclusion, relationship inference on a parental pair or on one of their offspring provides complementary information. When possible, sampling trios should be encouraged as it could allow spanning a wider range of potential relationships.

1. Introduction

Inferring the relationship that exists between the two partners in a couple is of interest for many areas of research from anthropology to genetics. It is informative of the mating habits and marriage patterns in a given population and allows comparative studies between populations (Romeo and Bittles, 2014). Several such studies have been performed in different human and animal populations based on pedigree records (see for example a recent work by Zlotogora and Shalev, 2014 in a Muslim village) or population surveys of the number of marriages between relatives based on church records (Sutter and Goux, 1964).

Recent advances in molecular genetics have made it possible to obtain genotype information for hundreds of thousands of markers spanning the whole genome. This genetic information can be used to estimate the kinship coefficients between pairs of individuals. This is now routinely done as a quality control step to identify related individuals in a sample and discard them to avoid false

* Correspondence to: Genetic Variation and Human Diseases Lab, Inserm U946, Fondation Jean Dausset-CEPH, 27 rue Juliette Dodu, 75010 Paris, France.
E-mail address: anne-louise.leutenegger@inserm.fr (A.-L. Leutenegger).

http://dx.doi.org/10.1016/j.tpb.2015.09.002

positives in case–control association studies (Voight and Pritchard, 2005). Unknown relatedness between individuals might then be discovered as it was, for example in the Hapmap data (Pemberton et al., 2010). When genetic information is available on spouses, it is then possible to get an overview of the realized relationships between them. Indeed, the pedigree only gives the expected relatedness and does not directly provide the true proportion of their genome that they really share (Speed and Balding, 2015). This was recently illustrated in both human and animal data (Colonna et al., 2007; Wang et al., 2014). Knowing this realized relationship might be of interest to identify regions of the genome that could harbor disease related genes.

Several different methods have been developed to estimate kinship coefficients between two individuals and infer their possible relationship or even reconstruct pedigrees from genetic data. These methods aim to identify regions of the genome that were inherited by the two individuals from a common ancestor and that are therefore identical-by-descent (IBD). They can be divided into two groups: single point methods that use the information at each marker independently and multipoint methods that take into account linkage between markers (see Browning and Browning, 2012 for a review). The latter methods have been shown to allow a better detection of distant relationships between individuals.

In parallel, similar methods have been developed to estimate inbreeding coefficients and identify genomic regions shared homozygous-by-descent (HBD) by a single individual (Leutenegger et al., 2003). These methods have mostly been used in the context of homozygosity mapping to identify genes involved in rare recessive monogenic diseases (Leutenegger et al., 2006) or genomic regions potentially harboring rare recessive variants involved in complex diseases (Génin et al., 2012). However, it is also possible to exploit the realized inbreeding in a population to learn about the mating habits in this population. Indeed, since the inbreeding coefficient of an individual is the same as their parents' kinship coefficient (Malécot, 1948), one can infer parental relationships using one of their offspring. We have recently proposed to do so with the individuals from the Human Genome Diversity Panel (HGDP-CEPH) to infer the mating habits of world-wide populations (Leutenegger et al., 2011). We have developed software that allows inferring the most likely relationship of the parents from the available genetic data of the offspring (Gazal et al., 2014a). Such indirect inference, based on offspring data, presents the advantage of being much simpler in terms of sampling than a direct inference from the parents. Indeed, sampling couples could be difficult and perhaps more prone to ascertainment bias than sampling isolated individuals. The cost is also double since two individuals need to be genotyped to estimate the kinship coefficient, compared to only one when estimating inbreeding. However, the available information in a couple is richer than in a single individual as we are then looking at four haplotypes instead of only two haplotypes. Finally, inference from a couple tells us about *potential* mating in the population but inference from an individual is informative about *realized* mating in the population.

Genome-based kinship and inbreeding coefficients are only equal in expectation with a large variability around this expected value (Donnelly, 1983; Leutenegger et al., 2003). To date however, no study has compared the estimates obtained from the genetic data on couples and from the genetic data on one of their offspring except for some studies that focused on assortative mating and aimed at identifying regions of the genome where offspring were either more or less similar than expected given the kinship of their parents (Laurent et al., 2012; Laurent and Chaix, 2012).

In this paper, we are interested in comparing the relationship inference obtained from the genetic data on either a pair of individuals or one of their offspring and the estimation of the proportion of genome shared IBD (kinship coefficient of the pair) or HBD (inbreeding coefficient of the offspring). To do so, we performed a simulation study on trio data with different parental relationships and compare (1) for the relationship inference, the results obtained using RELPAIR for a pair of individuals (Epstein et al., 2000) and using FSuite for a single individual, and (2) for the estimation of genome sharing proportions, the results of PLINK (Purcell et al., 2007), GIBDLD (Han and Abney, 2013) and FSuite (Gazal et al., 2014a).

2. Materials and methods

2.1. Estimating genomic kinship and inbreeding coefficients

Approaches to estimate the genomic kinship and inbreeding coefficients can be organized into three main categories. The first category of approaches rely on the allele frequencies at each marker considered independently (single-point). They can either use method of moments (MoM) estimation (Purcell et al., 2007; Yang et al., 2011) or maximum likelihood estimation (Thompson, 1975; Milligan, 2003; Polasek et al., 2010). The second category of approaches rely on the segmental nature of IBD (Purcell et al., 2007; Gusev et al., 2009). Finally, the third category of approaches that rely on both the marker allele frequencies and the segmental nature of IBD through hidden Markov models (HMM) (Leutenegger et al., 2003; Browning, 2008; Browning and Browning, 2010; Han and Abney, 2011; Brown et al., 2012; Han and Abney, 2013; Gazal et al., 2014a). Here, we focus on the single-point MoM approaches as implemented in PLINK (Purcell et al., 2007) and the multi-point approaches as implemented in GIBDLD (Han and Abney, 2013) and FSuite (Gazal et al., 2014a).

PLINK option—het allows the estimation of the genomic inbreeding coefficient F_{PLINK} as the genome-wide excess homozygosity. It is obtained as a function of the number of observed homozygous loci and the allele frequencies.

PLINK option—genome allows the estimation of the genomic kinship coefficient K_{PLINK}. PLINK provides both $\hat{\pi}$, which is twice the kinship coefficient, and the probabilities k_i of sharing i alleles IBD between two individuals, with the following relation between these different quantities: $K_{PLINK} = \hat{\pi}/2 = 0.5*k_2+0.25*k_1$. Note that when neither individual in the pair is inbred, k_i probabilities are also referred to as Cotterman's k coefficients (Cotterman, 1940). The k's are a function of the number of loci with 0, 1 or 2 alleles identical-by-state and the allele frequencies.

For the multi-point approaches, FSuite provides the maximum likelihood estimate (MLE) of the genomic inbreeding coefficient F_{FSUITE}. Let X_k denote the HBD state (i.e., $X_k = 1$ if the 2 alleles at locus k within the individual are IBD, 0 otherwise), and Y_k the genotype of the individual at locus k ($k = 1$ to N the total number of loci). The HBD process of an individual is approximated by a Markov chain, the transition probabilities $P(X_k|X_{k-1})$ depending on F the inbreeding coefficient, A the rate of HBD state change per cM and t_k the genetic distance between adjacent loci. The model requires the specification of the transition probabilities between the different HBD states at adjacent markers. These different transition probabilities are given in Leutenegger et al. (2003). For example, the probability for staying HBD at marker k given HBD at marker $k-1$ is: $P(X_k = 1|X_{k-1} = 1) = \left(1 - e^{-At_k}\right) F + e^{-At_k}$. The model also requires the specification of emission probabilities $P(Y_k|X_k)$ that depend on the allele frequencies at locus k. These allele frequencies can be estimated on the studied sample, or on a reference panel (such as HGDP-CEPH or HapMap panels) if the studied sample is too small to estimate them. Parameters F and A are then estimated by maximum likelihood.

GIBDLD for the estimation of the genomic kinship coefficient K_{GIBDLD} between two individuals relies on a similar model. The observed data Y_k are the unphased genotypes of the two

individuals at locus k. The hidden variable S_k has 9 possible values which are the 9 IBD states for the pair of individuals (Harris, 1964; Jacquard, 1970). The software provides an estimate of the genomic kinship coefficient derived from the posterior probabilities of the IBD states:

$$\begin{aligned} K_{GIBDLD} &= \frac{1}{N}\sum_{k=1}^{N}\Big\{P\left(S_k=1|\underline{Y}\right)+\frac{1}{2}\Big(P\left(S_k=3|\underline{Y}\right) \\ &\quad +P\left(S_k=5|\underline{Y}\right)+P\left(S_k=7|\underline{Y}\right)\Big)+\frac{1}{4}P\left(S_k=8|\underline{Y}\right)\Big\} \\ &= \frac{1}{N}\sum_{k=1}^{N}\Big\{\underline{\Delta}_1^{(k)}+\frac{1}{2}\left(\underline{\Delta}_3^{(k)}+\underline{\Delta}_5^{(k)}+\underline{\Delta}_7^{(k)}\right)+\frac{1}{4}\underline{\Delta}_8^{(k)}\Big\} \end{aligned}$$

where $\underline{\Delta}_i^{(k)}=P\left(S_k=i|\underline{Y}\right)$ is the posterior probability of IBD state i at marker k, and $\underline{Y}=(Y_1,Y_2,\ldots,Y_N)^t$ denotes the genotypes of the two individuals at all markers.

When neither individual in the pair is inbred, which will be the case for all the scenarios considered in this study, the previous formula reduces to the formula used in PLINK:

$$\begin{aligned} K_{GIBDLD} &= \frac{1}{N}\sum_{k=1}^{N}\Big\{\frac{1}{2}P\left(S_k=7|\underline{Y}\right)+\frac{1}{4}P\left(S_k=8|\underline{Y}\right)\Big\} \\ &= \frac{1}{N}\sum_{k=1}^{N}\Big\{\frac{1}{2}P\left(X_k=2|\underline{Y}\right)+\frac{1}{4}P\left(X_k=1|\underline{Y}\right)\Big\} \\ &= \frac{1}{N}\sum_{k=1}^{N}\Big\{\frac{1}{2}\underline{k}_2^{(k)}+\frac{1}{4}\underline{k}_1^{(k)}\Big\} \end{aligned}$$

where X_k is here the number of IBD alleles between the 2 individuals, and $\underline{k}_i^{(k)}$ is the posterior probability of sharing i IBD alleles at marker k. In this particular case, the transition probabilities for sharing 1 allele IBD in markers $k-1$ and k can be approximated similarly to FSuite by $P\left(X_k=1|X_{k-1}=1\right)=\left(1-e^{-At_k}\right)k_1+e^{-At_k}$ where A is here the rate of IBD state change per cM (Han and Abney, 2011). As in FSuite, parameters k_1 and A are estimated by maximum likelihood.

Note that some of the methods described above have been developed to take into account linkage disequilibrium (LD) that is present in dense SNP chip data and that can bias kinship/inbreeding estimates (Browning, 2008). PLINK methods are described by their authors as not LD-sensitive, even if it is advised to perform some LD pruning on the data before running the calculations (Anderson et al., 2010). The HMM implemented in FSuite needs markers to be in minimal LD and therefore FSuite generates several (default is 100) random submaps with 1 marker per 0.5 cM, and reports the median of the F estimations. Using 1 marker every 0.5 cM was found to provide accurate inbreeding estimation when running simulations based on real human haplotypes (Gazal et al., 2014b). Finally, GIBDLD HMM models LD on a sample by conditioning the emission probabilities on the 20 previous markers through a linear model.

2.2. Relationship inference

Relationship inference can be performed by comparing the likelihoods of several relationships and selecting the relationship of highest likelihood.

For a single individual, FSuite provides the likelihood for a given relationship of the parents by using the (F, A) expected from the pedigree. F is obtained by a standard path-counting approach (Wright, 1922) and A as $m/(100(1-F))$, where m is the number of meiosis in the inbreeding loop (Leutenegger et al., 2003; Thompson, 2013). In FSuite version 1.0.3, the relationships considered are avuncular (AV) $=(0.125, 0.057)$, double first cousin $(2\times 1C)=(0.125, 0.069)$, first cousin $(1C)=(0.0625, 0.064)$, second cousin $(2C)=(0.015625, 0.081)$, and unrelated $(UNR)=(0.001, 0.001)$.

For a pair of individuals, RELPAIR (Boehnke and Cox, 1997; Epstein et al., 2000) also uses a HMM to provide the likelihood for a given relationship. The observed data are the unphased genotypes of the two individuals while the hidden data are the number of IBD alleles (0, 1 or 2). The transition probabilities are derived specifically for each relationship by using the exact 2-locus IBD probabilities. The relationships considered in the latest version of the software (2.0.1) are: monozygotic twins (MZ), parent–offspring (PO), full-sibs (FS), half-sibs (HS), grand-parent grand-child (GG), AV, 1C, and UNR.

2.3. Simulation study

To compare the accuracies of the different strategies described above, we simulated genetic data for trios with non-inbred parents related through 10 different relationships: FS, HS, GG, AV, $2\times 1C$, 1C, quadruple second cousin (cyclic type, $4\times 2C$), 2C, third cousin (3C) and UNR. For each relationship, 1000 trios were simulated with the same strategy as in Gazal et al. (2014b). Briefly, we first simulated the recombination process for the whole genealogy with MORGAN2.9's Genedrop program. Genedrop assigns to each pedigree founder two labels, and simulates the recombination process between the two homologous copies of each chromosome based on the genetic distances between markers. Based on the pedigree founder labels, it is possible to determine the IBD/HBD states and compute the realized K of the parents and the realized F of the child of each trio as the proportion of genome in cM being IBD and HBD, respectively. To have realistic LD patterns, pedigree founder haplotypes were drawn without replacement among the 5412 WTCCC control haplotypes of 517,291 autosomal SNPs. Allele frequencies were computed from these haplotypes, and were used as inputs for PLINK, FSuite, RELPAIR. Note that even if true haplotypes were available from our simulations, none of the compared methods used phase information.

Single-point estimates were obtained from PLINK version 1.90b2b on a subset of 87,130 pruned SNPs. This subset of SNPs was obtained following Anderson et al. (2010)'s guidelines, by performing a heavy pruning (PLINK—indep-pairwise 50 5 0.20 option) on genotype data of the 2706 unrelated WTCCC control individuals.

Multi-point kinship estimates were obtained by running the GIBDLD option of IBDLD version 3.2. In order to remove markers with similar genotypes (i.e. very high LD ($r2 > 0.90$)), a light pruning (PLINK option—indep-pairwise 50 5 0.90) selecting 321,927 markers was first performed. Then, LD was modeled on the 5412 WTCCC control haplotypes through the IBDLD—phased option. Multi-point inbreeding estimates were obtained by running FSuite version 1.0.3 with default options, i.e. on 100 submaps selecting 1 marker every 0.5 cM (around 6500 markers per submap).

As the HMM implemented in RELPAIR also needs markers to be in minimal LD, RELPAIR was run on the same submaps as FSuite. The relationship inferred the largest number of times on the 100 submaps was reported. Here, RELPAIR never inferred a pair of individuals as MZ or PO. In order to make inferences from RELPAIR and FSuite comparable, we modified FSuite to only infer relationships from the same set of relationships used by RELPAIR. Namely, we modified FSuite by considering also FS, HS and GG offspring, and by discarding $2\times 1C$ and 2C offspring. For FS offspring, we computed likelihoods with $(F, A)=(1/4, 4/75)=(0.25, 0.053)$, while we used $(F, A)=(1/8, 8/175)=(0.125, 0.046)$ for HS/GG offspring. As done with RELPAIR, the relationship inferred the largest number of times on the 100 submaps was reported for FSuite.

Table 1
One-locus and two-locus IBD and HBD probabilities for avuncular (AV), half-sib (HS) and grand-parent/grand-child (GG).

		HS	GG	AV
1-locus	$K = F$	0.125	0.125	0.125
	k_0	0.5	0.5	0.5
	k_1	0.5	0.5	0.5
	k_2	0	0	0
2-locus[a]	$k11 =$ IBD1	0.5ψ	$0.5(1-\theta)$	$0.5[(1-\theta)\psi + 0.5\theta]$
	$f11 =$ HBD1	$0.125(1-\theta)^2\psi$	$0.125(1-\theta)^2\psi$	$0.125(1-\theta)^2[0.5\theta + \psi(1-\theta)]$

[a] θ = recombination fraction between the 2 loci. $\psi = \theta^2 + (1-\theta)^2$.

Because the realized IBD can vary substantially for a given relationship (Leutenegger et al., 2003; Hill and Weir, 2011), we did not compare the estimated coefficients with the ones expected from the genealogy. Rather, we estimated accuracies of kinship and inbreeding coefficients by measuring, for each trio, the difference between estimated and realized kinship coefficients (δK), and the difference between estimated and realized inbreeding coefficients (δF), respectively. For each simulated relationship, we measured the bias, the standard deviation (sd), and the root mean square error (RMSE) of the $1000\delta K$ and δF with RMSE $= \sqrt{bias^2 + sd^2}$.

3. Results

3.1. Theoretical comparison

FSuite, GIBDLD and RELPAIR all rely on a HMM to model the observed genotypes. Thanks to this multi-point modeling, these approaches can differentiate between relationships that have otherwise identical single-locus IBD probabilities and that single-point approaches (e.g. PLINK) cannot therefore differentiate (Thompson, 1986, 1988).

In the specific cases of GG, HS, AV, the kinship and inbreeding coefficients expected from the genealogy are the same ($1/8 = 0.125$) and so also are the k coefficients $(k_0, k_1, k_2) = (1/2, 1/2, 0)$. The 2-locus IBD 1 probability ($k11$) are not the same when $\theta \neq 0$ whereas offspring of HS or GG pairs have identical 2-locus HBD 1 probability $(f11) = 0.125(1-\theta)^2[\theta^2 + (1-\theta)^2]$ (Table 1). These two relationships are hence not differentiable based on the genetic data on a single offspring (even when using a multi-point approach) whereas they are differentiable based on the genetic data on the parental pair.

In order to compute the likelihood of each relationship, RELPAIR uses internally the genealogy specific $k11$ probabilities presented in Table 1. This is the reason why in RELPAIR the number of possible relationships is limited. On the contrary, GIBDLD and FSuite model the 2-locus IBD or HBD probabilities as a function of the one-locus IBD (resp. HBD) probability $k1$ (resp. F). The transition probabilities are assumed to be the same whatever the genealogy and depend on 2 parameters (A and k coefficients for GIBDLD, and A and F for FSuite, see Method section) that are estimated by maximum likelihood. This provides more flexibility as it does not require the pre-specification of the possible genealogies but it comes at the cost of an approximation in the 2 IBD probabilities. This can be seen in Fig. 1, which compares the estimated $k11$ (resp. $f11$) of GIBDLD (res. FSuite) with the true ones (Table 1). These approximations however are slightly better for the IBD obtained by GIBDLD than for the HBD obtained from FSuite. The other interest of the plot is to illustrate how much we can differentiate GG, HS, AV with parental pairs and much less with one offspring.

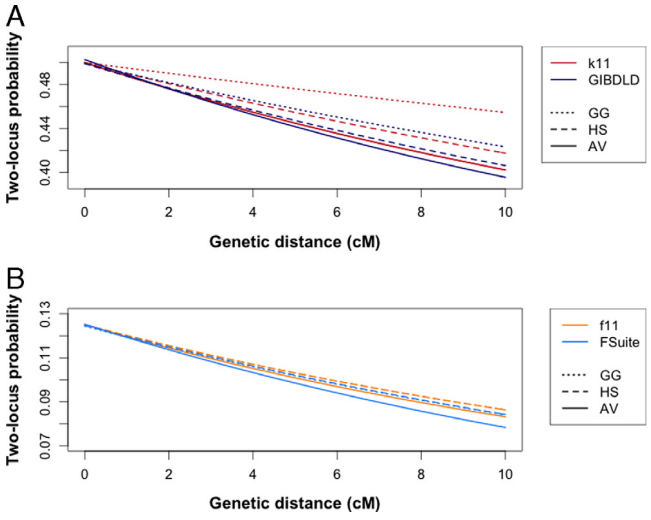

Fig. 1. Comparison between true and estimated two-locus probabilities for being IBD = 1 or HBD = 1 at both loci. (A) Comparison of the true probability for being IBD at two loci ($k11$) and its estimation by GIBDLD. (B) Comparison of the true probability for being HBD at two loci ($f11$) and its estimation by FSuite. Half sibs (HS), avuncular (AV) and grand-parent/grand-child (GG). In Figure (B), lines for HS and GG overlap. Note that this figure differs from the one presented in Thompson (1986), because it is given as a function of the genetic distance (over a short distance, 10 cM maximum) rather than recombination fraction. The genetic distances were derived from the recombination fractions through Haldane mapping function. The k_1, A (GIBDLD), F and A (FSuite) values necessary for the computation of the 2-locus probabilities are fixed to the values obtained on the true IBD or HBD data from the simulation study (Table S1).

3.2. Genomic kinship and inbreeding coefficients

3.2.1. Realized genomic kinship and inbreeding coefficients

First, one can see that there is a much higher variability around the value expected from the genealogy for the inbreeding coefficient (F) than for kinship coefficient (K) (Fig. 2). As previously described by Hill and Weir (2011), this is a consequence of Mendelian sampling and could be explained by the fact that from the parents to the offspring, there are two additional meioses. For instance, for 1C, the 95% variation interval around 0.0625 is [0.0198, 0.1113] for the former and [0.0425, 0.0843] for the latter.

Second, for more remote relationships, not all trios show relatedness (Table 2). For 2C, $K > 0$ in all 1000 trios but $F = 0$ in 16 trios. For 3C, 735 trios have $F > 0$ and $K > 0$, 249 trios have $K > 0$ and $F = 0$, and 16 have $F = 0$ and $K = 0$. Note that trios with $K = 0$ or $F = 0$ will not be taken into account to compute relationship inference rates.

3.2.2. Estimated genomic kinship and inbreeding coefficients

For each trio, we estimated the K of the parents and the F of the child by the single-point methods implemented in PLINK, and by the multi-point methods implemented in GIBDLD and FSuite. For both types of methods, we also observe a much higher variability around the value expected from the genealogy for F than

Table 2
Relationship inference rate with RELPAIR and FSuite. Each line gives results for 1000 simulated trios. The second column gives the number of simulated trios where parents share no IBD segment (realized $K = 0$), and the number of simulated trios where the child has no HBD segment (realized $F = 0$). For example, among the 1000 simulated 3C trios, 735 trios have $F > 0$ and $K > 0$, 249 trios have $K > 0$ and $F = 0$, and 16 have $F = 0$ and $K = 0$. These pairs/offspring have been removed to compute the relationships inference rates of 2C and 3C. Rates in bold give the most inferred relationship. See Fig. 1 legend for names of relationships.

	$(K = 0, F = 0)$	RELPAIR						FSuite				
		FS	HS	GG	AV	1C	UNR	FS	HS/GG	AV	1C	UNR
FS	(0, 0)	**1.000**	–	–	–	–	–	**0.968**	0.006	0.026	–	–
HS	(0, 0)	–	**0.772**	0.075	0.152	0.002	–	0.027	**0.596**	0.225	0.153	–
GG	(0, 0)	–	0.024	**0.976**	–	–	–	0.026	**0.597**	0.218	0.159	–
AV	(0, 0)	–	0.410	0.003	**0.581**	0.006	–	0.043	0.287	**0.579**	0.092	–
2 × 1C	(0, 0)	**0.983**	0.002	–	0.012	0.004	–	0.066	0.062	**0.805**	0.067	–
1C	(0, 0)	–	0.006	–	0.005	**0.990**	–	–	0.064	0.069	**0.866**	0.002
4 × 2C	(0, 0)	0.012	–	–	0.001	**0.987**	–	–	0.011	0.134	**0.854**	0.001
2C	(0, 16)	–	–	–	–	**0.654**	0.346	–	–	–	**0.752**	0.248
3C	(16, 265)	–	–	–	–	0.018	**0.982**	–	–	–	0.323	**0.677**
UNR	(1000, 1000)	–	–	–	–	–	**1.000**	–	–	–	–	**1.000**

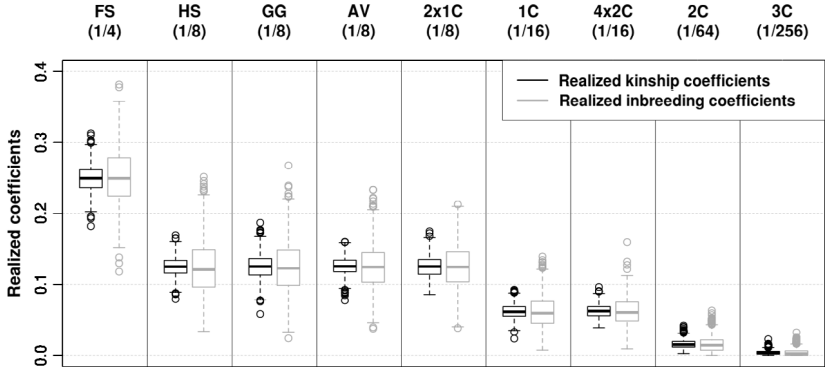

Fig. 2. Realized kinship coefficients (K) and inbreeding coefficients (F). This figure shows the boxplots of realized K and F for the 1000 trios of each relationship. Black boxplot represents realized K, while gray boxplot represents realized F. FS: Full Sibs, AV: Avuncular, HS: Half Sibs, GG: Grandparent–Grandchild, 2 × 1C: Double 1st cousins, 1C: 1st cousins, 4 × 2C: Quadruple 2nd cousins, 2: 2nd cousins, 3C: 3rd cousins. Numbers between brackets are the coefficients expected from the genealogy.

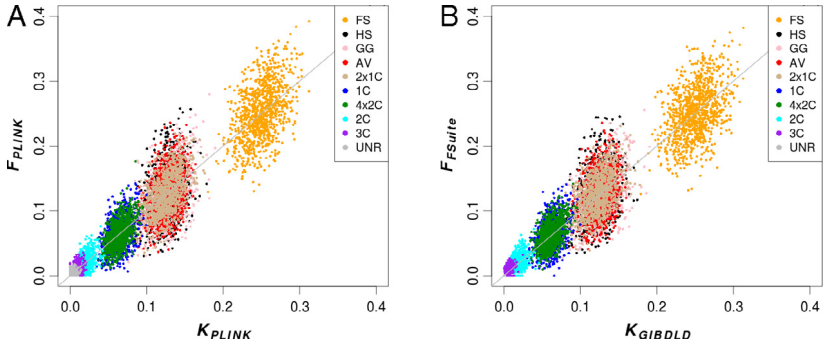

Fig. 3. Comparison of kinship and inbreeding coefficient estimates. Figure (A) compares single-point estimates (K_{PLINK} and F_{PLINK}). Figure (B) compares multi-point estimates (K_{GIBDLD} and F_{FSuite}). See Fig. 1 legend for names of relationships.

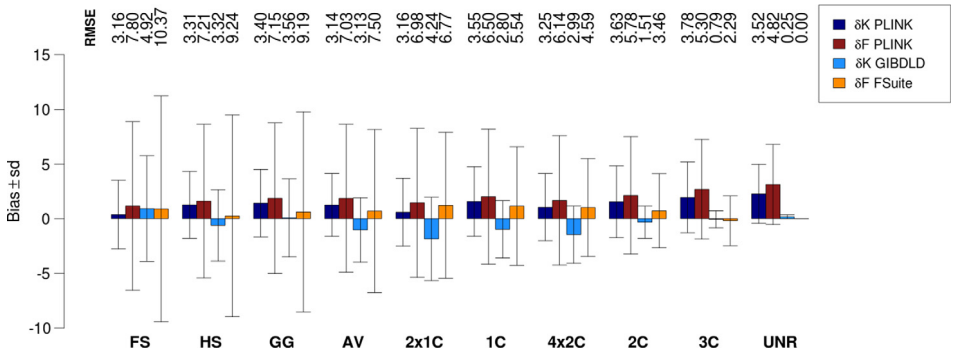

Fig. 4. Accuracy of the different estimators for various relationships. This figure shows the bias (in bar) of the difference between the estimated and realized coefficients (δK and δF) with ±1 standard deviation (sd, whole lines). Numbers on top are root mean square errors (RMSE). Scale of this figure is 10^{-3}. See Fig. 1 legend for names of relationships.

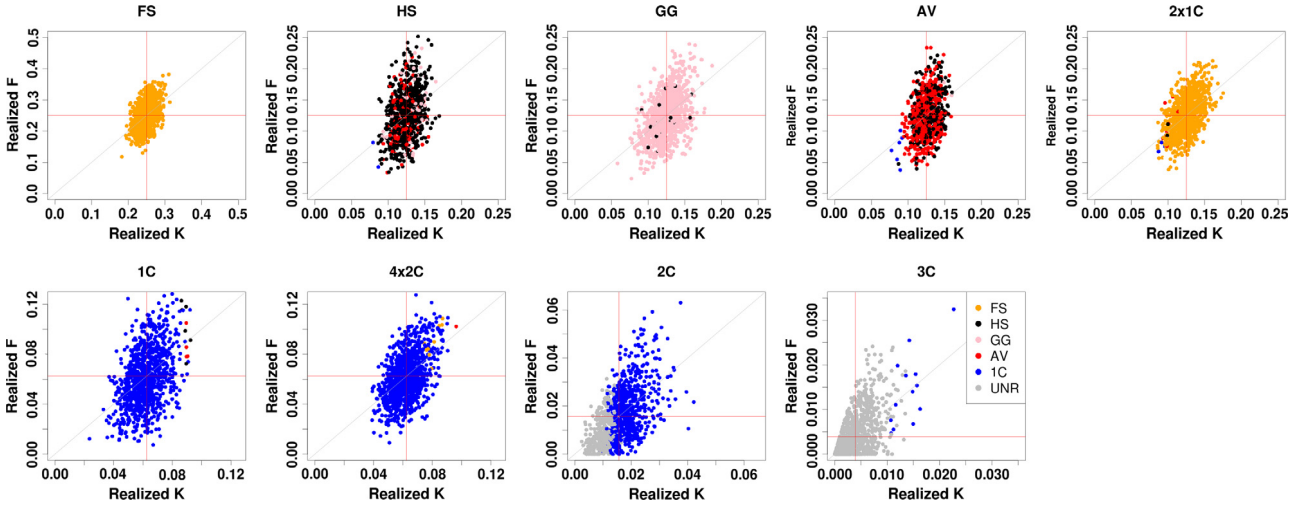

Fig. 5. Relationship inference with RELPAIR. Each dot represents a trio, with the realized kinship coefficient (K) of the parents on x-axis, and the realized inbreeding coefficient (F) of the child on y-axis. The colors give the relationship inferred the highest number of times by RELPAIR on the 100 submaps. Red lines represent the coefficients expected from the genealogy. See Fig. 1 legend for names of relationships. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

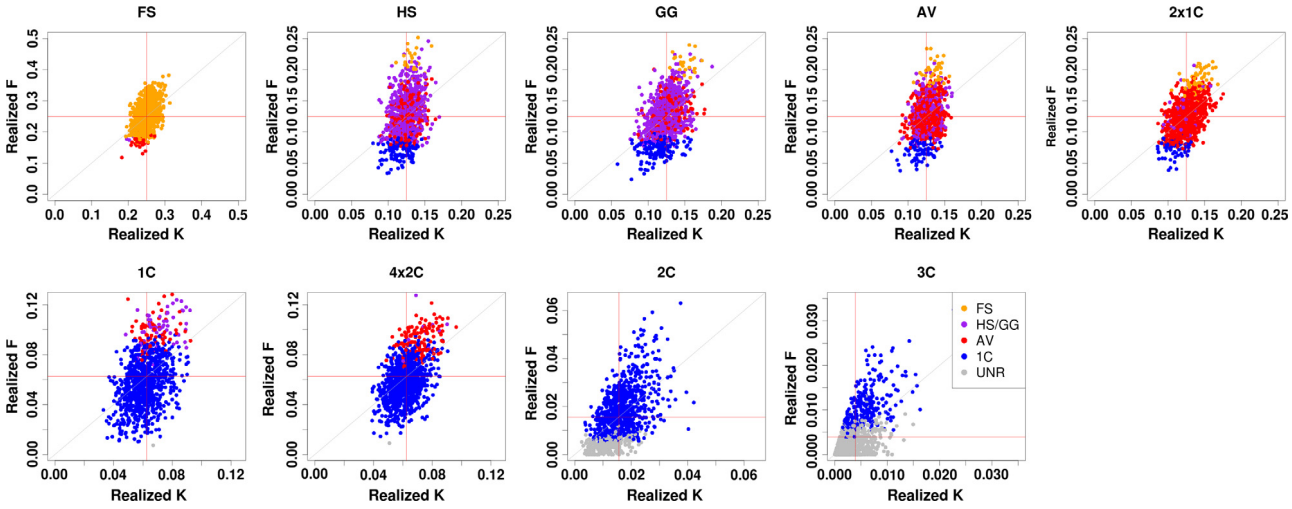

Fig. 6. Relationship inference with FSuite. Each dot represents a trio, with the realized kinship coefficient (K) of the parents on x-axis, and the realized inbreeding coefficient (F) of the child on y-axis. The colors give the relationship inferred the highest number of times by FSuite on the 100 submaps. Red lines represent the coefficients expected from the genealogy. See Fig. 1 legend for names of relationships. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for K (Figs. 3, S1 and S2). For single point estimates, we observe a constant variability (i.e. standard deviation) for δK whatever the relationship, while the variability of δF slowly decreases with more remote relationships (Fig. 4). For multipoint estimates, we observe that δK and δF variability decreases slowly and quickly, respectively, with the degree of relationship (Fig. 4). The RMSE of single-point estimates are equivalent or lower than the ones of multi-point estimates for close relationships (FS, HS, GG, AV and 2×1C), but are higher for other relationships. For 2C, 3C, and UNR, estimating F with FSuite is even more accurate than estimating K with PLINK.

3.3. Relationship inference from a pair of individuals or a single offspring

We then compared for each trio the accuracy of relationship inference with RELPAIR and FSuite (Figs. 5 and 6 and Table 2). For trios whose true relationships are tested by both methods (FS, HS, GG, AV and 1C), relationship inference rates of RELPAIR are always higher than those of FSuite. RELPAIR infers accurately 100% of FS, 97.6% of GG, and 99.0% of 1C, while it has more difficulty in distinguishing HS and AV with relationship inference rates of 77.2% and 58.1%, respectively. FSuite has high inference rates for FS and 1C offspring (96.8% and 86.6%, respectively), but has low ability to distinguish offspring whose expected inbreeding coefficient is 1/8 (59.6%, 59.7% and 57.9% for HS, GG and AV respectively).

For trios whose true relationships are not tested by both methods, FSuite gives the most coherent results. For 2×1C offspring, they are mainly inferred as AV offspring by FSuite (80.5%), which is consistent with their expected F (1/8), while pairs of 2×1C are mainly inferred as FS ($K = 1/4$) by RELPAIR (98.3%). This last result is due to the fact that 2×1C pairs can share both their chromosomes in some regions of the genome (IBD state $S_k = 7$ or k_2), and that FS is the most remote relationship in RELPAIR allowing for such situation. Finally, for more remote simulated relationships (2C, 3C), even when we removed pairs with no IBD segment, RELPAIR infers most pairs sharing at least one IBD segment as

unrelated (34.6% and 98.2%, respectively). On the contrary, FSuite was able to infer the genetically inbred offspring of these relationships as 1C offspring (75.2% and 32.3% for 2C and 3C, respectively) while correctly inferring all UNR offspring as outbred (Tables 2 and S2).

Note that when we add back to FSuite the possibility to test $2 \times 1C$ and 2C relationships, we observe that FSuite accurately infers 65.4% of $2 \times 1C$ and 81.7% of 2C (Table S2 and Figure S3). The rates of inbred 2C and 3C inferred as outbred decrease from 24.8% and 67.7% to 11.6% and 44.8%, respectively.

4. Discussion

We have compared the relationship inference and genome sharing coefficients obtained from the genetic data of either a pair of individuals or of one of their offspring using both single-point and multi-point approaches.

For close relationships (large shared segments), single-point approaches were found to perform better or similarly to multi-point methods at estimating kinship and inbreeding coefficients. But for more remote relationships (1C or less), multi-point methods gave the best estimates. Indeed, the length of IBD and HBD segments decreases with the number of meioses, and it is easier to detect small HBD segments, where there are no haplotype-phase uncertainty, than to detect small IBD segments (Browning and Browning, 2010).

When comparing the estimates obtained on the pairs of parents against those obtained on one of their offspring, we found, as expected, that there was always more variability for the single individual than for the pair of individuals. The genomic kinship and the genomic inbreeding coefficients were correlated but their relationship could be far from the bisecting line, as the HBD proportion of the genome of an individual can deviate from the IBD proportion of the parents due to Mendelian sampling (Fig. 5). This was observed both on the realized and estimated coefficients. About the large variability of FSuite, here in order to use the same marker sets between FSuite and RELPAIR, which is limited to 10,000 loci, we took the option of creating random submaps of markers spaced 0.05 cM apart. We would now recommend relying on the recombination hotspots (McVean et al., 2004; Winckler et al., 2005) to construct these random maps that will then contain more markers but too many for RELPAIR. Using this fine-scale information on recombination would allow a decrease in the variability of the results obtained with FSuite (Gazal et al., 2014b).

For the relationship inference, we only considered multi-point approaches. Again we found a difference for close and remote relationships. For close relationships, having the parental pair was always better than relying on one of their offspring. For more remote relationships however, we were able to detect the presence of inbreeding in one of the offspring when we could not detect kinship in the parental pair. This was especially the case for second and third cousin relationships. This is interesting as reported pedigrees will be less reliable for these remote relationships.

We found in our study that estimates obtained based on the two haplotypes of a single individual have a larger variance than those obtained based on the four haplotypes of the parents. This result was expected given the fact that the realized inbreeding coefficient for a given relationship has also a larger variance than the realized kinship coefficient. It is important to account for that difference when comparing kinship and inbreeding coefficient estimates as in studies of assortative mating that aims at finding regions of the genome where these coefficients differ.

In pedigree reconstruction such as PRIMUS (Staples et al., 2014) the single point IBD estimations are used. We have seen here that this is fine for close relationships but it could be interesting to evaluate the approach using multi-point estimates to help on some more remote relationships.

In all the scenarios studied here, we have considered only outbred parents. It could be of interest to study the properties of the methods when parents are themselves inbred. Indeed, in populations where marriages between relatives are encouraged, it is often the case that parents are both inbred and related. This would have an impact on the IBD sharing probabilities as shown in Génin and Clerget-Darpoux (1996, 1998), Génin et al. (1998) and in Liu and Weir (2004, 2005) when considering affected sib-pairs. Based, on these results, we can expect that single-point methods such as PLINK that rely on the 3 IBD states would overestimate the kinship coefficient whereas methods such as GIBDL that model the 9 identity states will correctly estimate it. Interestingly, in this case, estimates derived from the genetic information on the offspring will not be affected as only inbreeding coefficients of common ancestors of the two parents have some impact on the inbreeding coefficient of the descendants. Some more work however is needed to study how inbreeding of the parents will impact relationship inference.

In conclusion, relationship inference on a parental pair or on one of their offspring provides complementary information and allows the spanning of a wider range of potential relationships. Of course, trio sampling is more complicated than either single individual or individual couples. But when possible, it opens new possibilities of inference. To take full advantage of such trio design, methodological development accounting for identity by descent sharing between and within the 3 individuals would however be required.

Acknowledgments

We thank Mourad Sahbatou (Fondation Jean Dausset-CEPH) for his help with RELPAIR. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.tpb.2015.09.002.

References

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., Zondervan, K.T., 2010. Data quality control in genetic case-control association studies. Nat. Protoc. 5, 1564–1573. http://dx.doi.org/10.1038/nprot.2010.116.

Boehnke, M., Cox, N.J., 1997. Accurate inference of relationships in sib-pair linkage studies. Am. J. Hum. Genet. 61, 423–429. http://dx.doi.org/10.1086/514862.

Brown, M.D., Glazner, C.G., Zheng, C., Thompson, E.A., 2012. Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics 190, 1447–1460. http://dx.doi.org/10.1534/genetics.111.137570.

Browning, S.R., 2008. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics 178, 2123–2132. http://dx.doi.org/10.1534/genetics.107.084624.

Browning, S.R., Browning, B.L., 2010. High-resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. 86, 526–539. http://dx.doi.org/10.1016/j.ajhg.2010.02.021.

Browning, S.R., Browning, B.L., 2012. Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet. 46, 617–633. http://dx.doi.org/10.1146/annurev-genet-110711-155534.

Colonna, V., Nutile, T., Astore, M., Guardiola, O., Antoniol, G., Ciullo, M., Persico, M.G., 2007. Campora: a young genetic isolate in South Italy. Hum. Hered. 64, 123–135. http://dx.doi.org/10.1159/000101964.

Cotterman, C.W., 1940. Calculus for statistico-genetics (Ph.D. Thesis), Ohio State University, Published in "Genetics and Social Structure", P.A. Ballonoff ed., Academic Press, New York, 1974.

Donnelly, K.P., 1983. The probability that related individuals share some section of genome identical by descent. Theor. Popul. Biol. 23, 34–63.

Epstein, M.P., Duren, W.L., Boehnke, M., 2000. Improved inference of relationship for pairs of individuals. Am. J. Hum. Genet. 67, 1219–1231. http://dx.doi.org/10.1016/S0002-9297(07)62952-8.

Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E., Leutenegger, A.-L., 2014a. FSuite: exploiting inbreeding in dense SNP chip and exome data. Bioinform. Oxf. Engl. 30, 1940–1941. http://dx.doi.org/10.1093/bioinformatics/btu149.

Gazal, S., Sahbatou, M., Perdry, H., Letort, S., Génin, E., Leutenegger, A.-L., 2014b. Inbreeding coefficient estimation with dense SNP data: comparison of strategies and application to HapMap III. Hum. Hered. 77, 49–62. http://dx.doi.org/10.1159/000358224.

Génin, E., Clerget-Darpoux, F., 1996. Consanguinity and the sib-pair method: an approach using identity by descent between and within individuals. Am. J. Hum. Genet. 59, 1149–1162.

Génin, E., Clerget-Darpoux, F., 1998. Reply to weeks and sinsheimer. Am. J. Hum. Genet. 62, 731–736. http://dx.doi.org/10.1086/301744.

Génin, E., Quesneville, H., Clerget-Darpoux, F., 1998. On the probability of identity states in permutable populations: reply to Cannings. Am. J. Hum. Genet. 62, 726–727.

Génin, E., Sahbatou, M., Gazal, S., Babron, M.-C., Perdry, H., Leutenegger, A.-L., 2012. Could inbred cases identified in GWAS data succeed in detecting rare recessive variants where affected sib-pairs have failed? Hum. Hered. 74, 142–152. http://dx.doi.org/10.1159/000346790.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., Pe'er, I., 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19, 318–326. http://dx.doi.org/10.1101/gr.081398.108.

Han, L., Abney, M., 2011. Identity by descent estimation with dense genome-wide genotype data. Genet. Epidemiol. 35, 557–567. http://dx.doi.org/10.1002/gepi.20606.

Han, L., Abney, M., 2013. Using identity by descent estimation with dense genotype data to detect positive selection. Eur. J. Hum. Genet. EJHG 21, 205–211. http://dx.doi.org/10.1038/ejhg.2012.148.

Harris, D.L., 1964. Genotypic covariances between inbred relatives. Genetics 50, 1319–1348.

Hill, W.G., Weir, B.S., 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93, 47–64. http://dx.doi.org/10.1017/S0016672310000480.

Jacquard, A., 1970. Structure Génétique des Populations. Masson & Cie. Ed., Paris.

Laurent, R., Chaix, R., 2012. MHC-dependent mate choice in humans: why genomic patterns from the HapMap European American dataset support the hypothesis. BioEssays News Rev. Mol. Cell. Dev. Biol. 34, 267–271. http://dx.doi.org/10.1002/bies.201100150.

Laurent, R., Toupance, B., Chaix, R., 2012. Non-random mate choice in humans: insights from a genome scan. Mol. Ecol. 21, 587–596. http://dx.doi.org/10.1111/j.1365-294X.2011.05376.x.

Leutenegger, A.-L., Labalme, A., Genin, E., Toutain, A., Steichen, E., Clerget-Darpoux, F., Edery, P., 2006. Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. Am. J. Hum. Genet. 79, 62–66. http://dx.doi.org/10.1086/504640.

Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemainque, A., Clerget-Darpoux, F., Thompson, E.A., 2003. Estimation of the inbreeding coefficient through use of genomic data. Am. J. Hum. Genet. 73, 516.

Leutenegger, A.-L., Sahbatou, M., Gazal, S., Cann, H., Génin, E., 2011. Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? Eur. J. Hum. Genet. EJHG 19, 583–587. http://dx.doi.org/10.1038/ejhg.2010.205.

Liu, W., Weir, B.S., 2004. Affected sib pair tests in inbred populations. Ann. Hum. Genet. 68, 606–619. http://dx.doi.org/10.1046/j.1529-8817.2004.00121.x.

Liu, W., Weir, B.S., 2005. Genotypic probabilities for pairs of inbred relatives. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1379–1385. http://dx.doi.org/10.1098/rstb.2005.1677.

Malécot, G., 1948. Les Mathématiques de l'Hérédité. Masson. Ed., Paris.

McVean, G.A., et al., 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304, 581–584.

Milligan, B.G., 2003. Maximum-likelihood estimation of relatedness. Genetics 163, 1153–1167.

Pemberton, T.J., Wang, C., Li, J.Z., Rosenberg, N.A., 2010. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. Am. J. Hum. Genet. 87, 457–464. http://dx.doi.org/10.1016/j.ajhg.2010.08.014.

Polasek, O., Hayward, C., Bellenguez, C., Vitart, V., Kolcić, I., McQuillan, R., Saftić, V., Gyllensten, U., Wilson, J.F., Rudan, I., Wright, A.F., Campbell, H., Leutenegger, A.-L., 2010. Comparative assessment of methods for estimating individual genome-wide homozygosity-by-descent from human genomic data. BMC Genomics 11, 139. http://dx.doi.org/10.1186/1471-2164-11-139.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. http://dx.doi.org/10.1086/519795.

Romeo, G., Bittles, A.H., 2014. Consanguinity in the contemporary world. Hum. Hered. 77, 6–9. http://dx.doi.org/10.1159/000363352.

Speed, D., Balding, D.J., 2015. Relatedness in the post-genomic era: is it still useful? Nat. Rev. Genet. 16, 33–44. http://dx.doi.org/10.1038/nrg3821.

Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., Nickerson, D.A., Below, J.E., 2014. PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am. J. Hum. Genet. 95, 553–564. http://dx.doi.org/10.1016/j.ajhg.2014.10.005.

Sutter, J., Goux, J.M., 1964. Decline of consanguineous marriages in France from 1926 to 1958. Eugen. Q. 11, 127–140.

Thompson, E.A., 1975. The estimation of pairwise relationships. Ann. Hum. Genet. 39, 173–188.

Thompson, E.A., 1986. Pedigree Analysis in Human Genetics. Johns Hopkins University Press Ed., Baltimore, MD.

Thompson, E.A., 1988. Two-locus and three-locus gene identity by descent in pedigrees. IMA J. Math. Appl. Med. Biol. 5, 261–279.

Thompson, E.A., 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194, 301–326. http://dx.doi.org/10.1534/genetics.112.148825.

Voight, B.F., Pritchard, J.K., 2005. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 1, e32. http://dx.doi.org/10.1371/journal.pgen.0010032.

Wang, H., Misztal, I., Legarra, A., 2014. Differences between genomic-based and pedigree-based relationships in a chicken population, as a function of quality control and pedigree links among individuals. J. Anim. Breed. Genet. Z. Für Tierz. Zücht. 131, 445–451. http://dx.doi.org/10.1111/jbg.12109.

Winckler, W., et al., 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308, 107–111.

Wright, S., 1922. Coefficient of inbreeding and relationship. Am. Nat. 56, 330–338.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82. http://dx.doi.org/10.1016/j.ajhg.2010.11.011.

Zlotogora, J., Shalev, S.A., 2014. Marriage patterns and reproductive decision-making in the inhabitants of a single Muslim village during a 50-year period. Hum. Hered. 77, 10–15. http://dx.doi.org/10.1159/000357945.